

Segmentation of multiple series using a Lasso strategy

Karine Bertin ^{*}; Xavier Collilieux [†]; Emilie Lebarbier [‡] and Cristian Meza [§]

Abstract

We propose a new semi-parametric approach to the joint segmentation of multiple series corrupted by a functional part. This problem appears in particular in geodesy where GPS permanent station coordinate series are affected by undocumented artificial abrupt changes and additionally show prominent periodic variations. Detecting and estimating them are crucial, since those series are used to determine averaged reference coordinates in geosciences and to infer small tectonic motions induced by climate change. We propose an iterative procedure based on Dynamic Programming for the segmentation part and Lasso estimators for the functional part. Our Lasso procedure, based on the dictionary approach, allows us to both estimate smooth functions and functions with local irregularity, which permits more flexibility than previous proposed methods. This yields to a better estimation of the bias part and improvements in the segmentation. The performance of our method is assessed using simulated and real data. In particular, we apply our method to data from four GPS stations in Yarragadee, Australia. Our estimation procedure results to be a reliable tool to assess series in terms of change detection and periodic variations estimation giving an interpretable estimation of the functional part of the model in terms of known functions.

1 Introduction

The objective of segmentation methods is to detect abrupt changes (called breakpoints) in a signal. Such segmentation problems arise in many areas: in biology for the detection of chromosomal aberrations (Picard et al., 2005; Lai et al., 2005), in meteorology and climate (Caussinus and Mestre, 2004) to homogenize temperature and precipitation series or in geodesy for the detection of changes in GPS location series (Williams, 2003). In the latter

^{*}CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile

[†]IGN LAREG, Université Paris Diderot Sorbonne Paris Cité, Paris, France ; Observatoire de Paris, SYRTE, CNRS, UPMC, Paris, France

[‡]AgroParisTech UMR518, Paris 5e and INRA UMR518, Paris 5e, FRANCE

[§]CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile

example, none of the currently used segmentation methods have been shown to perform best than time series visual inspection (Gazeaux et al., 2013). One of the motivations of this paper is to develop an automatic method to tackle the analysis of such type of data.

GPS permanent stations that continuously monitor their coordinates have been deployed all over the world for more than 20 years. Their three-dimensional coordinate series are usually post-processed by scientists from raw code and phase observations at a daily or weekly basis, yielding series up to 1000 or 7000 records with a typical precision of a few millimeters. Such series are used to determine accurate station velocities for tectonic and Earth’s mantle studies, with a typical magnitude of a few millimeters per year to about ten centimeters per year (King et al., 2010). Such long-term coordinates (mean positions and velocities) of a worldwide network of stations also materialize a Terrestrial Reference Frame which is used for mapping purposes or for studying slowly varying physical phenomena including sea level variations (Altamimi et al., 2007). In addition, coordinate time series themselves were analyzed to infer information about ice melting and climate change (Wu et al., 2012; Wu et al., 2011; Wahr et al., 2013).

The observed coordinate variations reflect the ground deformations at the station including tectonic signals (generally a trend, mostly in the horizontal components) as well as environmental signals from the vicinity of the station, such as soil moisture or atmospheric pressure changes. The latter could be approximated by periodic signals with dominant annual and semi-annual periods (Dong et al., 2002). The observational noise exhibits more autocorrelation at long periods (Williams et al., 2004) but specific systematic errors of small magnitudes also show up at some well known periods, which are either submultiple of 350.5 days (Ray et al., 2008) or annual like thermal deformation of the station monumentation and the ground. Abrupt changes from a few millimeters to meters are superimposed to those variations. They are related to instrumental changes (documented or not), GPS multiple signal reflection, earthquakes or changes in the raw data processing strategy. The detection of these offsets but also of the periodic components is fundamental for the above mentioned applications. Up to now, offsets are first identified visually and the periodic components are estimated in a second run for interpretation (van Dam et al., 2012).

It is common to observe the same situation in genomics for the detection of chromosomal aberrations since the biological phenomenon (corresponding to the segmentation) can be contaminated by a probe effect or a wave-effect (see Picard et al., 2011 and references therein). Neglecting these effects could generate false detection and leads to wrong conclusions about the aberrations. As illustrated in Section 4, other examples can be found where a set of biases represented by some functions needs to be adjusted within a segmentation model.

In all these data, the form of the functional biases are not always well specified, or are even unknown. Using a non-parametric approach is very useful since it does not require specification of the form of the functions to estimate. In this sense, Picard et al. (2011) proposed a semi-parametric approach to the joint segmentation of multiple series in the genomic application field. When the segmentation is specific to each series and the biases (probe effect or wave-effect) are shared by all the series, considering multiple series allows them to better estimate these biases and so to improve the segmentation. The model they proposed is split into two parts: a parametric part corresponding to the segmentation and a non-parametric part (the functional one) which is estimated using wavelets, splines or is viewed as a fixed effect. On the one hand, the estimation with spline or wavelet gives good results when the biases are smooth functions but fails when these present local irregularities. On the other hand, the approach with the fixed effect model tends to catch both local irregularities of the bias and the noise, which can produce erratic estimation of the bias part.

In this article, we propose a more flexible modelization of the functional part by estimating it using a dictionary approach. In other words, it is estimated by linear combinations of functions with different regularities: smooth functions (for example spline functions or Fourier functions) and more irregular functions (for example spiky functions). To select the relevant functions, we use a Lasso-type strategy introduced by Tibshirani (1996) and recently applied in a semi-parametric framework by Arribas-Gil et al. (2014) resulting in an estimation procedure with good practical and theoretical performance (oracle-type estimator). Lasso non-parametric estimators have several advantages. A first one is that the size of the dictionary can be large and this does not affect the computational cost of the method. As a consequence, many different functions can be put in the dictionary. This method is then very flexible and allows us to estimate functions with both smooth components and local irregularities. Moreover the resulting estimators are sparse linear combinations of the functions of the dictionary. In practice this is helpful for the interpretation of the results.

As usual in the segmentation context with a maximum penalized likelihood estimation framework, we first estimate the segmentation parameters and the non-parametric part, the number of segments being fixed. Then we apply a model selection strategy to choose this number. For the first task, the two parts can not be estimated simultaneously. Indeed in order to infer the breakpoint parameters, it is now well known that Dynamic Programming (DP) strategies remain among the most efficient. However this algorithm can only be applied when the contrast to be optimized is additive with respect to the segments (Bai and Perron, 2003; Caussinus and Mestre, 2004; Picard et al., 2005). And this is not the case when there is a global

parameter, as the function in our model (Bai and Perron, 2003). This is why, following Picard et al. (2011) or Bai and Perron (2003), our method consists in an iterative two-steps procedure which alternates between the segmentation issue and a Lasso-type estimation of the functional part.

We apply this strategy to simulated data where the functional part is a mixture of smooth functions and irregular functions. We obtain good results for both segmentation and functional bias parts and, in particular, we outperform the methods of Picard et al. (2011) with wavelet, spline or fixed effect. Moreover we apply our method to GPS data from Australian stations. For these data, we find several breakpoints of interest. The estimated non-parametric part is found to be relevant since the obtained periodic functions have been suggested in previous studies. Their amplitudes and phases are more relevant for geophysical interpretation (see for example Dong et al., 2002; van Dam et al., 2012 for such an interpretation), since they have been simultaneously estimated all together and jointly with the segmentation part.

This article is organized as follows. In Section 2, we present the semi-parametric segmentation model for multiple series. In Section 3, we describe our two-step iterative procedure based on DP for segmentation part and Lasso dictionary approach for the functional part given a fixed number of segments, and the model selection strategy for choosing the number of segments. In Section 4, a simulation study is carried out to assess the performance of our method comparatively to other methods and we illustrate the improvements obtained for a real climatic data set. In Section 5, we apply our method to the geodetic data described above and a final conclusion is given in Section 6.

2 Semi-parametric model

We observe M series. We note $y_m(t)$ the observed signal of the series m at time t and we suppose that it satisfies for $m \in \{1, \dots, M\}$

$$y_m(t) = \mu_m(t) + f(x_m(t)) + e_m(t), \quad (1)$$

where $\mu_m(t) = \mu_k^m$ if $t \in I_k^m = (\tau_{k-1}^m, \tau_k^m]$, x_m represents possible covariates (the simple one is the time t), f is an unknown function to be estimated, τ_k^m is the k th breakpoint of the series m , μ_k^m is the mean of the series m on the segment I_k^m and the $e_m(t)$ are i.i.d centered Gaussian with variance σ^2 . We note K_m the number of segments of the m th series and $K = \sum_{m=1}^M K_m$ the total number of segments. Note that the segmentation is specific to each series.

For $m \in \{1, \dots, M\}$, the series m has n_m observations in the times t_{mi} , $i \in \{1, \dots, n_m\}$, so the total number of observations is $N = \sum_{m=1}^M n_m$ and

the model is

$$\begin{aligned} y_{mi} &= \mu_{mi} + f(x_{mi}) + e_{mi}, \\ \forall i \in \{1, \dots, n_m\}, m \in \{1, \dots, M\}, \end{aligned} \quad (2)$$

where $y_{mi} = y_m(t_{mi})$, $x_{mi} = x_m(t_{mi})$, $e_{mi} = e_m(t_{mi})$ and $\mu_{mi} = \mu_m(t_{mi})$. We define the vectors $y_m := (y_{mi})_i$, $x_m := (x_{mi})_i$, $e_m := (e_{mi})_i$ and $\mu_m := (\mu_k^m)_k$.

The parameters of the model are the means μ_k^m , the breakpoints τ_k^m , the function f , the variance σ^2 and the number of segments K .

3 Estimation procedure

As usual in the segmentation estimation framework, the parameters are estimated for a fixed number of segments K for which we propose here a DP-Lasso procedure, then K is chosen using a model selection strategy.

3.1 A DP-Lasso estimation procedure

Following Bai and Perron (2003), we propose an iterative procedure that alternates the segmentation part with the estimation of f . The function f corresponds to a bias part common to each series and our objective is to estimate it non-parametrically using a Lasso-type method based on a dictionary approach. More specifically, we consider a collection of functions $\phi = \{\phi_1, \dots, \phi_J\}$ and we propose to estimate f by a linear combination of the functions ϕ_j ,

$$f_\lambda = \sum_{j=1}^J \lambda_j \phi_j, \quad \lambda \in \mathbb{R}^J.$$

In order to write our estimation algorithm in a matricial form, we concatenate the means vectors μ_m in a vector $\boldsymbol{\mu}$ of size $K \times 1$. We denote by \mathbf{T} ($[N \times K]$) the incidence matrix of breakpoints $\mathbf{T} = \text{Bloc}[\mathbf{T}_m]$ with $\mathbf{T}_m = \text{Bloc}[\mathbf{1}_{n_k^m}]$ of size $([n_m \times K_m])$, and with $n_k^m = \tau_k^m - \tau_{k-1}^m$ the length of k -th segment for series m . $\mathbf{T}\boldsymbol{\mu}$ corresponds to the segmentation part. Moreover, we concatenate the vectors y_m and x_m in the $([N \times 1])$ vectors \mathbf{Y} and \mathbf{X} . We denote by F the $[N \times J]$ matrix $F = (f_{i,j})$ where $f_{i,j} = \phi_j(\mathbf{X}_i)$.

We denote by $\mathbf{T}\boldsymbol{\mu}^{(h)}$ the segmentation estimated parameters, $(\sigma^{(h)})^2$ the estimated variance, $\lambda^{(h)}$ the estimated coefficients of the function f , and $f^{(h)} = f_{\lambda^{(h)}}$ the estimated function f at iteration (h) . At iteration $(h+1)$, we get:

- given $\lambda^{(h)}$, the segmentation parameters $\mathbf{T}\boldsymbol{\mu}^{(h+1)}$ are estimated by:

$$\mathbf{T}\boldsymbol{\mu}^{(h+1)} = \underset{\mathbf{T}\boldsymbol{\mu}}{\text{argmin}} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - F\lambda^{(h)}\|^2,$$

where $\|\cdot\|$ stands for the L_2 norm in \mathbb{R}^N . The problem is then reduced to segment $\mathbf{Y} - F\lambda^{(h)}$ into K segments. In the case of joint segmentation, Picard et al. (2011) proposed a double-stage of DP which used the multiple structure and allows us to obtain the best segmentation of all the series into K segments in a more reasonable computational time compared to the classical DP.

- given $\mathbf{T}\boldsymbol{\mu}^{(h+1)}$ and $\sigma^{(h)}$, the function f is estimated using a Lasso-type strategy:

$$f^{(h+1)} = f_{\lambda^{(h+1)}}$$

where $\lambda^{(h+1)}$ minimizes

$$\|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}^{(h+1)} - F\lambda\|^2 + 2 \sum_{j=1}^J r_{N,j} |\lambda_j|,$$

where following Arribas-Gil et al. (2014),

$$r_{N,j} = \sigma^{(h)} \|\phi_j\|_N \sqrt{\gamma \log J} \text{ with } \gamma > 2 \text{ and } \|\phi_j\|_N = \sqrt{\sum_{l=1}^N \phi_j^2(X_l)}.$$

- given $\mathbf{T}\boldsymbol{\mu}^{(h+1)}$ and $f^{(h+1)}$, the variance σ^2 is estimated by

$$(\sigma^{(h+1)})^2 = \frac{1}{N} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}^{(h+1)} - F\lambda^{(h+1)}\|^2.$$

The algorithm stops when the difference between parameters of two successive iterations is smaller than ϵ (10^{-3} in practice).

The final estimators are denoted $\hat{\tau}_k^m$, $\hat{\mu}_k^m$, $\widehat{\mathbf{T}\boldsymbol{\mu}}$, $\hat{\sigma}^2$, $\hat{\lambda}$ and $\hat{f} = f_{\hat{\lambda}}$.

Remark 1. *From a theoretical point of view, the condition $\gamma > 2$ ensures that the resulting estimator of f has good properties (oracle performance, see Arribas-Gil et al., 2014). However, in cases in which the Lasso estimation is performed within an iterative procedure involving the estimation of other parameters than f , the value of γ may also influence the stability of the whole iterative procedure. Then, γ should be chosen as close as possible to 2 while allowing for the stability of the iterative algorithm.*

3.2 Model selection

The last issue is the choice of the number of segments K . We propose here to use the modified BIC criterion proposed by Zhang and Siegmund (2007)

and successfully adapted to the joint segmentation by Picard et al. (2011):

$$\begin{aligned} mBIC_{\text{JointSeg}}(K) &= \log \left[\Gamma \left(\frac{N - K + 1}{2} \right) \right] \\ &- \left(\frac{N - K + 1}{2} \right) \log SS_{\text{wg}} + \left[\frac{1}{2} - (K - M) \right] \log(N) \\ &- \frac{1}{2} \sum_{m=1}^M \sum_{k=1}^{k_m} \log(\hat{\tau}_k^m - \hat{\tau}_{k-1}^m), \end{aligned}$$

where $SS_{\text{wg}} = \|\mathbf{Y} - \widehat{\mathbf{T}\boldsymbol{\mu}} - F\hat{\lambda}\|^2$.

4 Study of the performance of the method

In order to assess the performance of our procedure, so-called here *Lasso*, in Section 4.1, we conduct the simulation study described below. We also propose to compare our method to the work of Picard et al. (2011), where either the function f is estimated using splines or f is viewed as a fixed effect depending on the time t , i.e. $f(t) = \beta_t$. We call these two approaches *Spline* and *Position* respectively and we perform them on the simulated data using the `cghseg` R package, in particular using the `multiseg` R function. For our procedure, we develop our own functions in R using the `lars` R package to perform the Lasso estimation of f . In addition in Section 4.2, we illustrate on a climatic data set the need to model correctly the function f in order to avoid false detection in the segmentation.

4.1 Simulation study

Simulation design. We consider the model (1) for series $m \in \{1, \dots, M\}$ at time t :

$$y_m(t) = \mu_m(t) + f(t) + e_m(t), \quad t = 1, \dots, n \quad (3)$$

where $e_m(t) \sim \mathcal{N}(0, \sigma^2)$ i.i.d. The length n of the series is fixed and equal to 100. We consider two different numbers of series: $M \in \{10, 50\}$, and five values for error variance: $\sigma^2 \in \{0.1, 0.2, 0.5, 1.0, 1.5\}$. For each series, the number of segments K follows a Poisson distribution with mean $\bar{K} = 3$ and their positions are uniformly distributed. The mean value within each segment alternates between 0 and a value in $\{-2, -1, +1, +2\}$ with probability $\{0.2, 0.3, 0.3, 0.2\}$ respectively. The function f is generated as a mixture of a sine function with three peaks (see Figure 1):

$$\begin{aligned} f(t) &= 0.3 \times \sin \left(2\pi \frac{t}{20} \right) + 0.5 \mathbb{I}_{t=0.1 \times n} \\ &\quad - \mathbb{I}_{t=0.5 \times n} + 2 \mathbb{I}_{t=0.6 \times n}. \end{aligned} \quad (4)$$

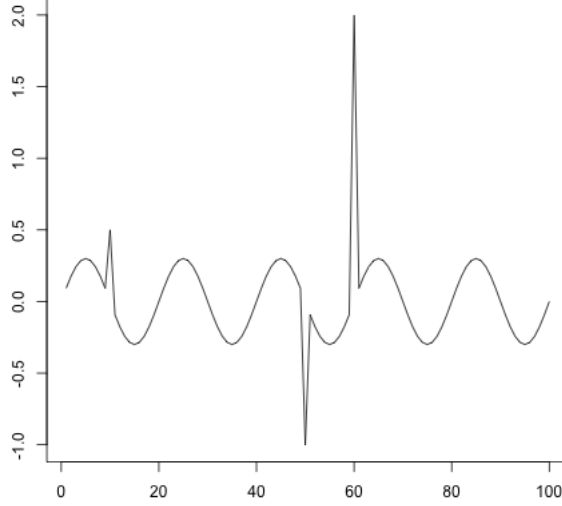


Figure 1: Simulated function f .

Each configuration, i.e. specific values of M and σ^2 , is simulated 100 times.

For the Lasso strategy, we use a dictionary with 150 functions: 128 Haar functions ($t \mapsto 2^{7/2} \mathbb{I}_{[0,1]} \left(\frac{2^7 t}{100} - k \right)$, $k = 0, \dots, 2^7 - 1$), the Fourier functions ($t \mapsto \sin \left(\frac{2\pi j t}{100} \right)$, $t \mapsto \cos \left(\frac{2\pi j t}{100} \right)$, $j = 1, \dots, 10$) and the functions $t \mapsto t$ and $t \mapsto t^2$. The Lasso estimator is obtained by LARS algorithm with $\gamma = 2.1$.

Quality criteria. To study the quality of the estimation, for each configuration, we consider several criteria:

- For the segmentation parameters, in order to study the global quality of the estimation, we consider the root-mean-square distance between the true mean and its estimate:

$$\text{RMSE}(\mu) = \left[\frac{1}{N} \sum_{m=1}^M \sum_{t=1}^n \{ \mu_m(t) - \hat{\mu}_m(t) \}^2 \right]^{1/2} \text{ where } N = M \times n.$$

Moreover, to study the performance of the estimation of the breakpoint positioning, we consider both the proportion of erroneously detected breakpoints among detected breakpoints (false discovery rate, FDR) and the proportion of undetected true breakpoints among true breakpoints (false negative rate, FNR).

- For the function f , the root-mean-square distance between f and its estimate:

$$\text{RMSE}(f) = \left[\frac{1}{n} \sum_{t=1}^n \left\{ f(t) - \hat{f}(t) \right\}^2 \right]^{1/2} \text{ is also considered.}$$

For each configuration, we consider the average of these criteria over the 100 simulations.

Comparison between *Lasso*, *Spline* and *Position*. Only the results with $M = 10$ are presented since the results for $M = 50$ leads to same conclusions.

Figure 2 presents the $\text{RMSE}(f)$ for the different methods with respect to σ . We observe that the larger is the noise, the worst is the estimation of f due to the confusion between the signal and the noise. Whatever the level of noise, *Lasso* outperforms *Position* and *Spline* in terms of the non-parametric part estimation. However, the behavior of *Position* and *Spline* is opposite with respect to σ . For small σ , *Spline* leads to bad performances compared to *Lasso* and *Position*. Indeed, as expected, *Spline* tends to capture the smooth part of the signal, i.e. the sinusoidal trend only, whereas the two others catch both the peaks and the trend. However, for large σ , it is more difficult to detect the peaks of the true function, resulting in closest results for *Lasso* and *Spline*. *Position* behaves worstly since, as mentioned in Picard et al. (2011), it tends to catch the trend but also the noise resulting in an erratic estimation of f . The bad estimation of f can have consequences on the segmentation estimation. Figure 3 summarizes the results for the segmentation estimation obtained with the different methods with respect to σ . In general, *Lasso* is slightly better than the two other methods. For σ larger than 0.5, the results are similar, even for *Position* for which f is not well estimated. However for small values of σ , since *Spline* does not detect the peaks, they are considered as breakpoints in the segmentation, leading to bad results: more segments are then detected (see $\hat{K} - K$), these false breakpoints then increase the FDR and so the $\text{RMSE}(\mu)$.

As a conclusion, *Position* and *Lasso* behave similarly for the estimation of the segmentation part. The main difference concerns the estimation of f which is less reliable for *Position*. An important advantage of *Lasso* is its flexibility in the sense that functions of different regularities can be included in the dictionary and in particular some functions chosen according to the knowledge of the expert. The final form of the estimator \hat{f} is a sparse linear combination of the dictionary functions that allows a possible interpretation of f compared to *Position* (see Section 5).

Discussion on the quality of the estimation with *Lasso*. We first compare the results obtained with the true and estimated number of segments. In Figure 2, we observe that the more difficult is the detection (more σ increases), more the number of segments is under-estimated. This result was expected and is now classical in the study of model selection for segmentation. Indeed, the number of segments is reduced in order to avoid false detection. This is illustrated by a less increase of the FDR obtained with the estimated number of segments compared to the true one (Figure 3). That leads to a better estimation in terms of segmentation (small $RMSE(\mu)$) and consequently to a better estimation of the function f (small $RMSE(f)$).

Segmentation and the estimation of f as a function of the number of series. Table 1 summarizes the relative differences for two criteria, the FDR and the root-mean-square of f between $M = 10$ and $M = 50$, for several values of σ as:

$$\begin{aligned} FDR^\sigma &= \frac{FDR_{10}^\sigma - FDR_{50}^\sigma}{FDR_{10}^\sigma}, \\ RMSE(f)^\sigma &= \frac{RMSE(f)_{10}^\sigma - RMSE(f)_{50}^\sigma}{RMSE(f)_{10}^\sigma}, \end{aligned}$$

where, for example, FDR_{10}^σ and $RMSE_{10}(f)^\sigma$ denote respectively the FDR and the root-mean-square of f for $M = 10$ series for a specific value of σ . Table 2 shows the percentage of the true functions of the simulated function f selected in the estimator \hat{f} against different values of σ , with $M = 10$ and $M = 50$ series. The ID function corresponds to the position of the true functions in the dictionary with size 150. Specifically, the first three functions (labels 13, 64 and 77) are Haar functions centered in 10, 50 and 60 and the function 137 is the function $x \mapsto \sin(2\pi\frac{5t}{100})$. In addition, a FDR criterion is calculated, corresponding to the number of false selected functions among the selected ones. As expected, the increase of the number of series improves the estimation of f (large $RMSE(f)^\sigma$). For small values of σ , the *Lasso* procedure leads to a good performance in terms of selected functions whatever the number of series: the number of selected functions is close to the true one, and among them all the true functions of the simulated function are retrieved with less false selection (small FDR function). That leads logically to an accurate estimation of f (small $RMSE(f)$ Figure 2). For noisy configurations (large σ), fewer functions are selected, which was expected. Indeed, in this case, there are more confusion between noise and signal, the small peaks (in particular ID 13 and 64) are more difficult to detect. This is particularly true for a small number of series. Remark that for $M = 50$, the ID 77 and 137 are always selected. Moreover, the better accuracy of the estimation of f observed for $M = 50$

Table 1: Comparison between $M = 10$ and $M = 50$ series for FDR and $RMSE(f)$ criteria.

σ	Relative differences	
	FDR^σ	$RMSE(f)^\sigma$
0.1	-	57.46
0.2	42.15	57.97
0.5	9.40	55.58
1.0	7.00	50.47
1.5	5.64	47.47

Table 2: Percentage, FDR and mean of number of functions selected by Lasso.

	σ	<i>ID function</i>				<i>FDR</i>	<i>Mean</i>
		13	64	77	137	function	length
M=10	0.1	100	100	100	100	0.052	4.27
	0.2	100	100	100	100	0.055	4.29
	0.5	26	99	100	100	0.064	3.53
	1.0	5	28	99	99	0.114	2.13
	1.5	0	12	73	76	0.137	1.9
M=50	0.1	100	100	100	100	0.059	4.31
	0.2	100	100	100	100	0.059	4.31
	0.5	100	100	100	100	0.068	4.36
	1.0	53	100	100	100	0.084	3.95
	1.5	18	92	100	100	0.108	3.6

leads to a better positioning of the breakpoints (see FDR^σ). This is less marked when σ is large.

4.2 Illustration

In this section, we want to illustrate the need to model correctly the function f in order to avoid false detection in the segmentation. To this end, we compare our procedure to the results obtained by Picard et al. (2011) in their study on harvest dates. In this application, the purpose is to detect changes in the agricultural practices by detecting changes in the grape harvest dates which are not due to the climatic effect. The data are harvest dates obtained at 10 French stations. The model they proposed is a mixed linear model containing a segmentation part, a random effect and a climatic effect modelled by a degree 2 polynomial according to the temperature. To compare with our proposed strategy, we avoid the random effect. The model

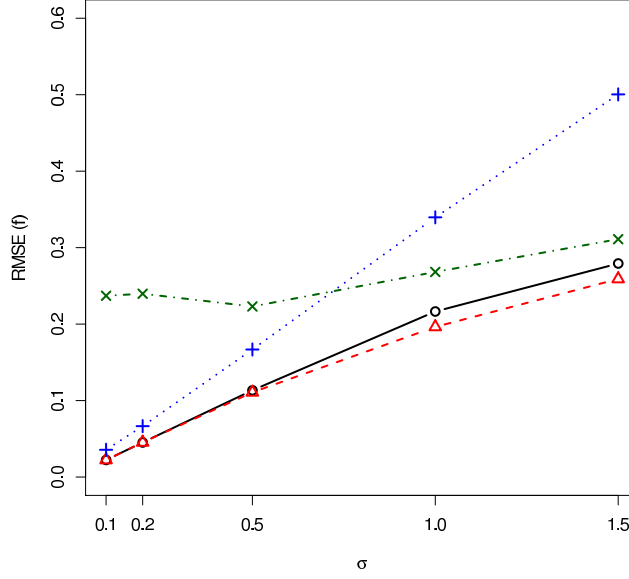


Figure 2: RMSE of f with respect to σ for *Lasso* \triangle , *Position* +, *Spline* \times and *Lasso* with the true number of segments \circ for $M = 10$.

is then written as follows:

$$y_m(t) = \mu_k^m + f(x_m(t)) + e_m(t), \text{ if } t \in I_k^m,$$

where $y_m(t)$ is the grape harvest date and $x_m(t)$ is the mean temperature of the year t for series m . In case (1), the form of the climatic effect f is fixed to be $f(x_m(t)) = bx_{mt} + cx_{mt}^2$. In case (2), no assumptions are made on the function f and it is estimated using our proposed procedure, for which we consider a dictionary with 36 functions compound with high resolution level Haar wavelets, Fourier basis, x , x^2 and x^3 . In the resulting estimator of f obtained with $\gamma = 2.1$, five functions are selected. Figure 4 represents the number of detected breakpoints per year over all the series for the two models. The result obtained in case (1) is slightly different from the one obtained in Picard et al. (2011). However, the most important difference compared to the result obtained by our proposed procedure concerns the year 2003 which corresponds to a very hot summer: that year is considered as a breakpoint in case (1) and not in case (2). This breakpoint appears in the series 6. Figure 5 represents respectively the harvest dates of the series 6 and its segmentation after correction in case (1) (segmentation of $y_t - \hat{b}x_t - \hat{c}x_t^2$). The temperature at year 2003 is 32.15. As shown in Figure 6, the correction of the harvest date at this year by $\hat{b}x_{mt} + \hat{c}x_{mt}^2$ is too strong compared to $\hat{f}(x_{mt})$ obtained in case (2) that is why a false breakpoint is added (see Figure 5 bottom).

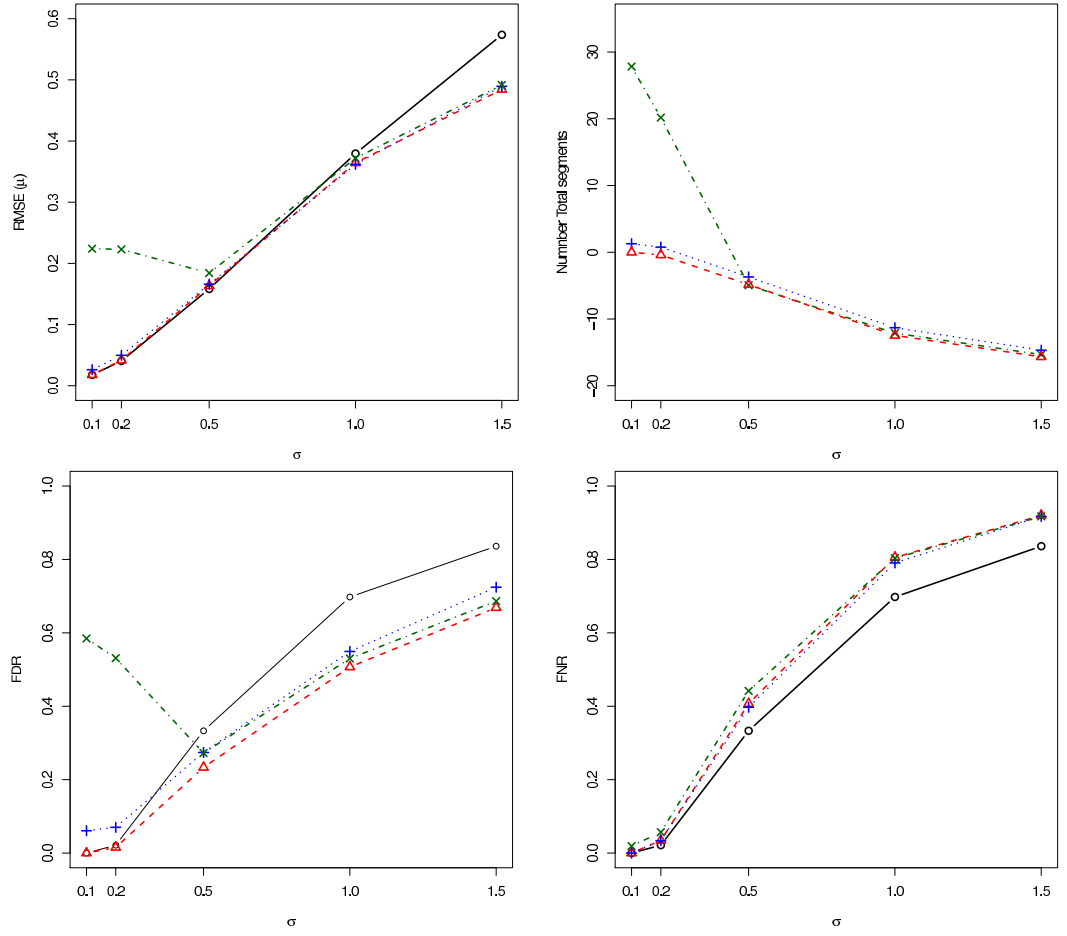


Figure 3: Results for $M = 10$ with respect to σ . Top: RMSE of μ on the left, $\hat{K} - K$ on the right. Bottom: FDR on the left and FNR on the right. *Lasso* \triangle , *Position* $+$, *Spline* \times and *Lasso* with the true number of segments \circ .

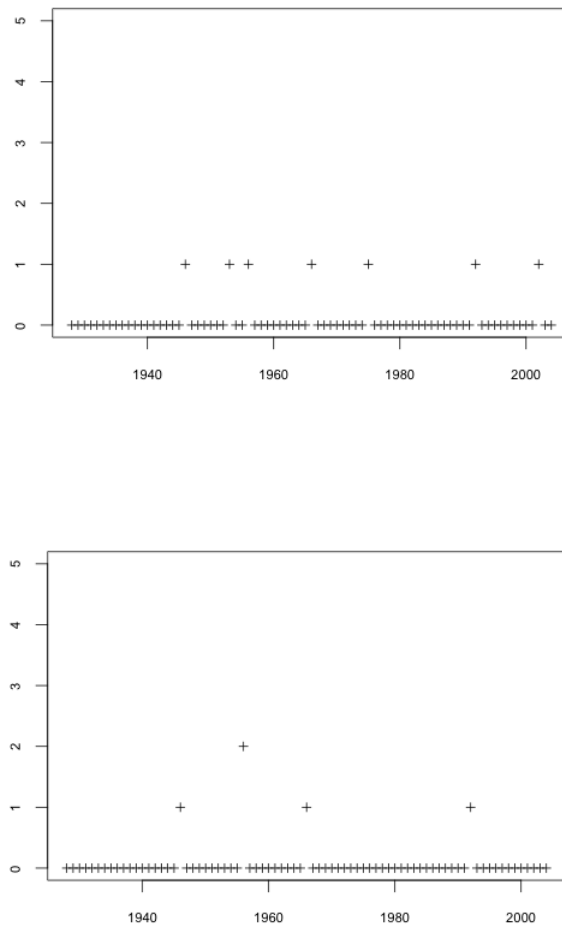


Figure 4: Number of the detected breakpoints over all the stations obtained in case (1) on the top and case (2) on the bottom.

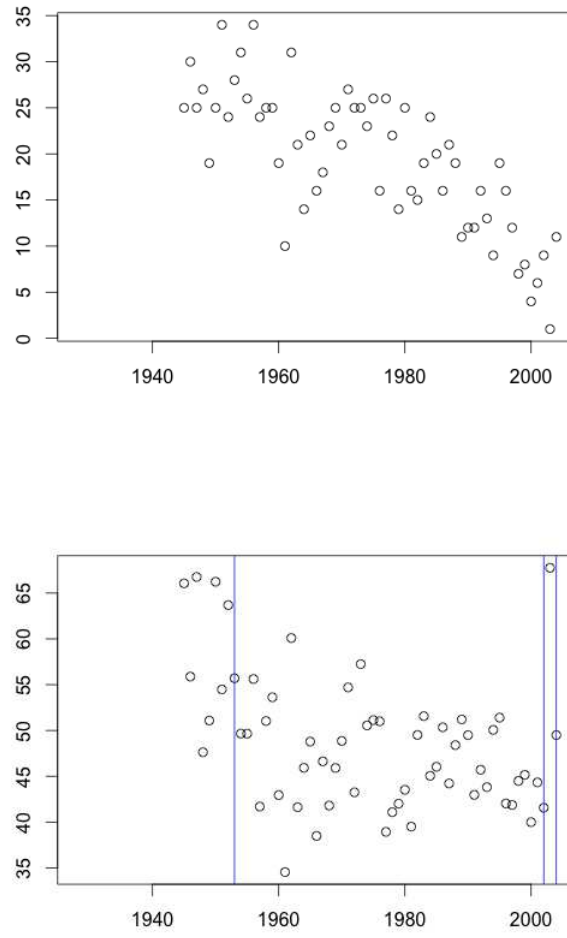


Figure 5: Top: harvest dates of the series 6. Bottom: obtained segmentation of the corrected series in case (1) (on $y_{mt} - \hat{b}x_{mt} - \hat{c}x_{mt}^2$).

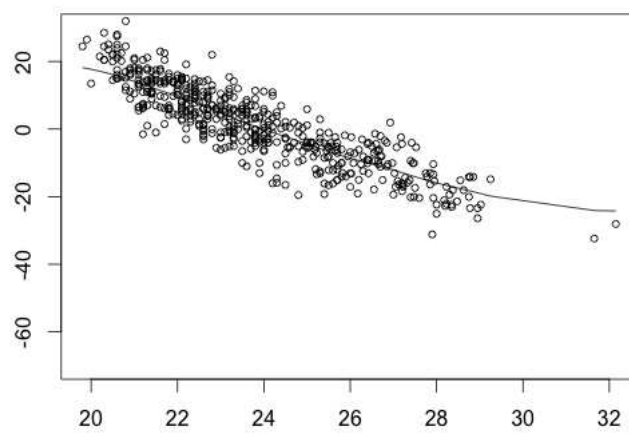
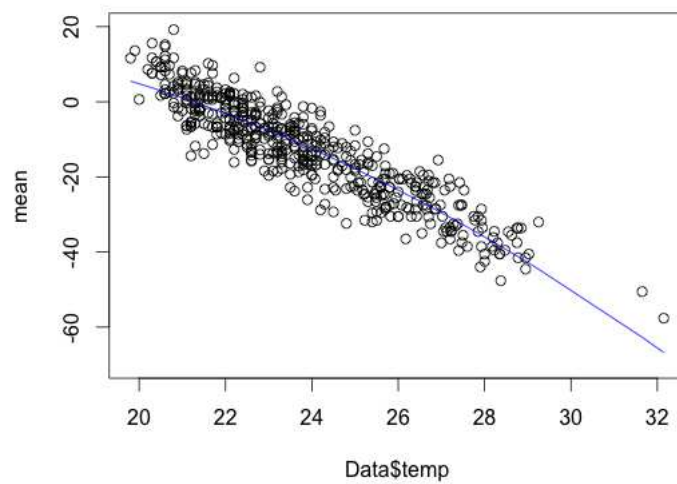


Figure 6: Fit of f in case (1) on the top and case (2) on the bottom.

5 Application

In this Section, we summarize the results obtained with our estimation procedure for the GPS dataset described in Introduction. In particular, we use the height coordinate series of four GPS stations in Australia located in Yarragadee (YAR1, YAR2, YAR3 and YARR). Those were computed by the Jet Propulsion Laboratory (JPL). They can be downloaded at ftp://sideshow.jpl.nasa.gov/pub/JPL_GPS_Timeseries/repro2011b/post/point/. We use the series from their first observations to the 22nd of June 2013 - series provided online are updated everyday. Then the model (1) is considered with $M = 4$ and $n_1 = 2862$, $n_2 = 5209$, $n_3 = 1443$ and $n_4 = 2197$, the respective lengths of the series. Here they have been averaged at weekly scale. For all these series, the ground motion is assumed to be identically observed and is described with function $f(t)$. Thus, equipment changes or malfunction at individual station should show up in the segmentation. For those series, JPL detected changes using a procedure based on sequential F-test applied to the tridimensional coordinate series (M. Heflin, personal communication, 2014).

We apply our proposed procedure to these series with a dictionary with 226 functions, which are only Fourier functions: $t \mapsto \sin(2\pi w_i t)$, $t \mapsto \cos(2\pi w_i t)$ where $w_i = i/T$, $T = \max(t) - \min(t)$ and T/i is larger than 8 weeks since smaller period amplitudes are generally negligible (see in Ray et al., 2008). Figure 7 shows the results for the four series: the obtained breakpoints in solid vertical lines, the known equipment changes in dashed vertical line and the estimated function f in solid line.

A total of 50 periods (62 bases) has been selected, among them the ones close to the well-known frequencies mentioned above (annual and semi-annual) and submultiples of the draconitic periods. 12 long periods - larger than 1 year - reflect well-known GPS low-frequency noise as already noticed by Amiri-Simkooei et al., 2007.

Height breakpoints are detected. Four (GPS week 1689 and 1707 of the series YAR2 and 1508 and 1559 of the series YAR3) correspond exactly to receiver and antenna changes. The changes at time 1205 of the series YAR2 is likely to be related to the equipment change at time 1166. In the same series, a change at time 1628 is detected. This change is not known from databases, however, it is also proposed by JPL. Compared to the JPL official list of changes, we found three additional changes for YAR2 at GPS week 1205 and the two validated changes at 1689 and 1707. Our two other additional changes at time 900 of the series YAR1 and at time 1191 of the series YARR are not reported by JPL. Up to now, no explanation has been supplied for those.

As a conclusion, our method found the same known breakpoints as JPL official list, but includes new validated one. Moreover the 62 bases function selected in the Lasso procedure furnish relevant geodetic information.

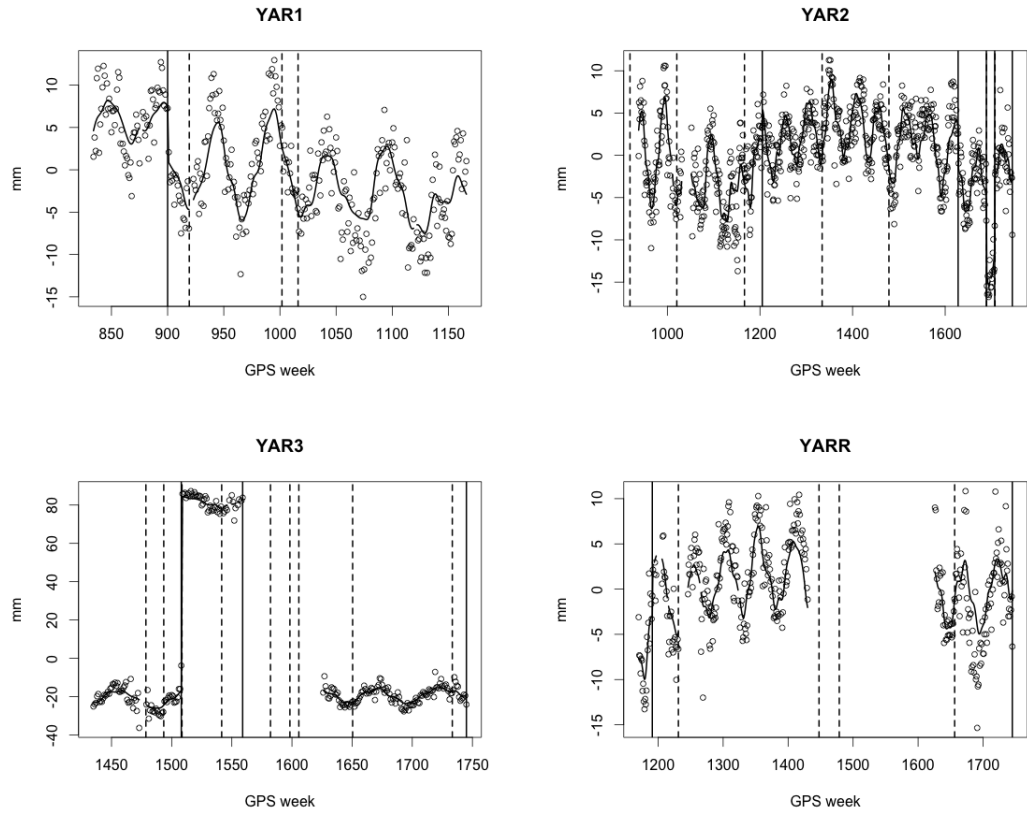


Figure 7: Results for height coordinate series of four GPS stations (YAR1, YAR2, YAR3 and YARR): obtained breakpoints in solid vertical lines; known equipment changes in dashed vertical line; estimated function f in solid line.

6 Conclusion

The proposed semi-parametric approach for the segmentation of single or multiple series has been shown to provide a valuable and reliable tool to assess changes and functional variations in series, as illustrated with our GPS height series. The search for functions that model ground motions and periodic errors here was crucial to provide the right segmentation of the series and reliable estimates of the breakpoint amplitudes. Conversely, because the segmentation is simultaneous and the number and location of the breakpoints unknown, estimated functions are also more reliable. They can be used to better interpret ground deformation observations or to enhance the piece-wise linear coordinate model of the Terrestrial Reference Frame (Petit and Luzum, 2010; Altamimi and Dermanis, 2012), widely used for geosciences and mapping applications. This would provide a significant improvement for the users since such coordinates are aimed to be extrapolated in the future (up to 5 years). Because the method is totally flexible and allows for a large number of functions to be included in the dictionary, it could also be applied to GPS series from active tectonic areas where the ground motion signal is more complex and should be modeled with additional functions.

Acknowledgements

Karine Bertin is supported by the grant ANILLO ACT-1112, CONICYT-PIA, Chile and FONDECYT project 1141258. Emilie Lebarbier is supported by the grant CONICYT 870100003 atracción de capital humano avanzado del extranjero. Cristian Meza is supported by the grant ANILLO ACT-1112, CONICYT-PIA, Chile and FONDECYT project 1141256.

References

- Altamimi, Z., X. Collilieux, J. Legrand, B. Garayt, and C. Boucher (2007). Itf2005: A new release of the international terrestrial reference frame based on time series of station positions and earth orientation parameters. *Journal of Geophysical Research* 112(B09401).
- Altamimi, Z. and A. Dermanis (2012). The choice of reference system in itrf formulation. In N. Sneeuw, P. Novák, M. Crespi, and F. Sansò (Eds.), *VII Hotine-Marussi Symposium on Mathematical Geodesy*, Volume 137 of *International Association of Geodesy Symposia*, pp. 329–334. Springer Berlin Heidelberg.
- Amiri-Simkooei, A. R., C. C. J. M. Tiberius, and P. J. G. Teunissen (2007).

- Assessment of noise in GPS coordinate time series: Methodology and results. *Journal of Geophysical Research (Solid Earth)* 112(B7).
- Arribas-Gil, A., B. K., M. C., and R. V. (2014). Lasso-type estimators for semiparametric nonlinear mixed-effects models estimation. *Statistics and Computing* 24(3), 443–460.
- Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *J. Appl. Econ.* 18, 1–22.
- Caussinus, H. and O. Mestre (2004). Detection and correction of artificial shifts in climate series. *Applied Statistics* 53, 405–425.
- Dong, D., P. P. Fang, Y. Bock, M. K. Cheng, and S. Miyazaki (2002). Anatomy of apparent seasonal variations from gps-derived site position time series. *Journal of Geophysical Research (Solid Earth)* 107(B4), ETG 9–1.
- Gazeaux, J., S. Williams, M. King, M. Bos, R. Dach, M. Deo, A. W. Moore, L. Ostini, E. Petrie, M. Roggero, F. N. Teferle, G. Olivares, and F. H. Webb (2013). Detecting offsets in gps time series: First results from the detection of offsets in gps experiment. *Journal of Geophysical Research: Solid Earth* 118(5), 2397–2407.
- King, M., Z. Altamimi, J. Boehm, M. Bos, R. Dach, P. Elsegui, F. Fund, M. Hernandez-Pajares, D. Lavallée, P. Cerveira, R. Riva, P. Steigenberger, T. van Dam, L. Vittuari, S. Williams, and P. Willis (2010). Improved constraints on models of glacial isostatic adjustment. *A review of the contribution of ground-based geodetic observations, Surveys in Geophysics* 31(5).
- Lai, W., M. Johnson, R. Kucherlapati, and P. J. Park (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21(19), 3763–3770.
- Petit, G. and B. Luzum (2010). Iers conventions (2010). (iers technical note ; 36). Technical report, Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie. inprint.
- Picard, F., E. Lebarbier, E. Budinska, and Robin (2011). Joint segmentation of multivariate gaussian processes using mixed linear models. *Comp. Stat. & Data Analysis* 55, 1160–1170.
- Picard, F., E. Lebarbier, M. Hoebeke, G. Rigai, B. Thiam, and S. Robin (2011). Joint segmentation, calling and normalization of multiple cgh profiles. *Biostatistics* 12(3), 413–428.
- Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin (2005). A statistical approach for CGH microarray data analysis. *BMC Bioinformatics* 6, 27.

- Ray, J., Z. Altamimi, X. Collilieux, and T. van Dam (2008). Anomalous harmonics in the spectra of GPS position estimates. *GPS Solutions* 12(1).
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288.
- van Dam, T., X. Collilieux, J. Wuite, Z. Altamimi, and J. Ray (2012). Non-tidal ocean loading: amplitudes and potential effects in gps height time series. *Journal of Geodesy* 86(11), 1043–1057.
- Wahr, J., S. A. Khan, T. van Dam, L. Liu, J. H. van Angelen, M. R. van den Broeke, and C. M. Meertens (2013). The use of gps horizontals for loading studies, with applications to northern california and southeast greenland. *Journal of Geophysical Research: Solid Earth* 118(4), 1795–1806.
- Williams, S. (2003). Offsets in global positioning system time series. *Journal of Geophysical Research (Solid Earth)* 108(19), 2310–+.
- Williams, S., Y. Bock, P. Fang, P. Jamason, R. Nikolaidis, L. Prawirodirdjo, M. M., and D. Johnson (2004). Error analysis of continuous GPS position time series. *Journal of Geophysical Research* 109(B18), B03412.
- Wu, X., X. Collilieux, Z. Altamimi, B. Vermeersen, R. Gross, and I. Fukumori (2012). Accuracy of the international terrestrial reference frame origin and earth expansion. *Geophysical Research Letters* 38(L13304).
- Wu, X., M. B. Heflin, H. Schotman, B. L. A. Vermeersen, D. Dong, R. S. Gross, E. R. Ivins, A. W. Moore, and S. E. Owen (2011). Simultaneous estimation of global present-day water transport and glacial isostatic adjustment. *Nature Geoscience* 3(9), 642–646.
- Zhang, N. R. and D. O. Siegmund (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63(1), 22–32.